pages 1-2 bridging paragraph:

The speech signal can be roughly divided into voiced and unvoiced regions. The voiced speech is periodic with a varying level of periodicity. The unvoiced speech does not display any apparent periodicity and has a noisy character. Transitions between voiced and unvoiced regions as well as temporary sound outbursts (e.g., plosives like "p" or "t") are neither periodic nor clearly noise-like. In low-bit rate speech coding, applying different techniques to various speech regions can result in increased efficiency and perceptually more accurate signal representation. In coders which use linear prediction, the linear LP-synthesis filter is used to generate output speech. The excitation of the LP-synthesis filter models the LP-analysis residual which maintains speech characteristics: it is periodic for voiced speech, noise-like for unvoiced segments, and neither for transitions or plosives. In the Code Excited Linear Prediction (CELP) coder, the LP excitation is generated as a sum of a pitch synthesis-filter output (sometimes implemented as an entry in an adaptive codebook) and an innovation sequence. The pitch-filter (adaptive codebook) models the periodicity of the voiced speech. The unvoiced segments are generated from a fixed codebook which contains stochastic vectors. The codebook entries are selected based on the error between input (target) signal and synthesized speech making CELP a waveform coder. T.Moriya and M.Honda "~~Seech~~ Speech Coder Using Phase Equalization and Vector Quantization", Proc. IEEE ICASSP 1701 (1986), describe a phase equalization filtering to take advantage of perceptual redundancy in slowly varying phase characteristics and thereby reduce the number of bits required for coding.

A1

pages 2-3 bridging paragraph:

In the Mixed Excitation Linear Prediction (MELP) coder, the LP excitation is encoded as a superposition of periodic and non-periodic components. The periodic part is generated from waveforms, each representing a pitch period, encoded in the frequency domain. The non-periodic part consists of noise generated based on signal

A2

correlations in individual frequency bands. The MELP-generated voiced excitation contains both (periodic and non-periodic) components while the unvoiced excitation is limited to the non-periodic component. The coder parameters are encoded based on an error between parameters extracted from input speech and parameters used to synthesize output speech making MELP a parametric coder. The MELP coder, like other parametric coders, is very good at reconstructing the strong periodicity of steady voiced regions. It is able to arrive at a good representation of a strongly periodic signal quickly and well adjusts to small variations present in the signal. It is, however, less effective at modeling ~~aperiodic~~ non-periodic speech segments like transitions, plosive sounds, and unvoiced regions. The CELP coder, on the other hand, by matching the target waveform directly, seems to do better than MELP at representing irregular features of speech. It is capable of maintaining strong signal periodicity but, at low bit-rates, it takes CELP longer to "build up" a good representation of periodic speech. The CELP coder is also less effective at matching small variations of strongly periodic signals.

---

page 3, first full paragraph:

These observations suggest that using both CELP and MELP (waveform and parametric) coders to a represent speech signal would provide many benefits as each coder seems to be better at representing different speech regions. The MELP coder might be most effectively used in periodic regions and the CELP coder might be best for unvoiced, transitions, and other ~~nonperiodic~~ non-periodic segments of speech. For example, D. L. Thomson and D. P. Prezas, "Selective Modeling of the LPC Residual During Unvoiced Frames; White Noise or Pulse Excitation," Proc. IEEE ICASSP, (Tokyo), 3087–3090 (1986) describes an LPC vocoder with a multipulse waveform coder, W. B. Kleijn, "Encoding Speech Using Prototype Waveforms," 1 IEEE Trans.Speech and Audio Proc., 386–399 (1993) describes a CELP coder with the Prototype Waveform Interpolation coder, and E. Shlomot, V. Cuperman, and A. Gersho, "Combined Harmonic and Waveform Coding of Speech at Low Bit Rates,"

Proc. IEEE ICASSP (Seattle), 585–588 (1998) describes a CELP coder with a sinusoidal coder.

A3

page 4, first full paragraph:

These ~~feature~~ features each has advantages including a low-bit-rate hybrid coder using the voicing of weakly-voiced frames to enhance the waveform coder and avoiding phase discontinuities at the switching between parametric and waveform coded frames.

A4

pages 14-15 bridging paragraph:

(11) filter the input speech frame into five frequency bands (0-500 Hz, 500-1000 Hz, 1000-2000 Hz, 2000-3000 Hz, and 3000-4000 Hz). For each frequency band again use the partitioning into six 44-sample subframes with each subframe having four pitch estimates as in the preceding ~~fpitch[]~~ fpitch[k] candidates derivation. Then for k= 0,1,2,3 and j=1,2,3,4,5 compute the j-th bandpass correlation bpcorr[j,k] as the sum over subframes of cross-correlations using the k-th pitch estimate (omitting any adjustment factor).

A5

page 18, second paragraph from bottom:

(2) extract a waveform from each residual by an N-point discrete Fourier transform. Note that the Fourier coefficients thus correspond to the amplitudes of the pitch frequency and its harmonics for the subframe. The gain parameter is the energy of the residual divided by N, which is just the average squared sample amplitude. Because the Fourier transform is complex symmetric (due to the speech being real), only the harmonics up to N/2 need be retained. Also, the dc (~~zeroth~~ zeroth harmonic) can be ignored.

A6

page 29, first paragraph:

To facilitate arbitrary switching between a waveform coder and a parametric coder, preferred embodiments may remove the phase component from the target signal for the waveform (CELP) coder. The target signal is used by the waveform coder in its signal analysis; by removing the phase component from the target, the preferred embodiments make the target signal more similar to the signal synthesized by the parametric coder, thereby limiting switching artifacts. Indeed, Figure 6a illustrates an example of a residual for a weakly-voiced frame in the ~~lefthand~~ left-hand portion and a residual for a strongly-voiced frame in the ~~righthand~~ right-hand portion. Figure 6b illustrates the removal of the phase components of the weakly-voiced residual, and the weakly-voiced residual now appears more similar to the strongly-voiced residual which also had its phase components removed by the use of amplitude-only Fourier coefficients. Recall that in the foregoing MELP description the waveform Fourier coefficients X[n] (DFT of the residual) was converted to amplitude-only coefficients |X[n]| for coding; and this conversion to amplitude-only sharpens the pulse in the time domain. Note that the alignment phase relates to the time synchronization of the synthesized pulse with the input speech. The zero-phase equalization for the CELP weakly-voiced frames performs a sharpening of the pulse analogous to that of the MELP's conversion to amplitude-only; the zero-phase equalization does not move the pulse and no further time synchronization is needed.

page 33, first full paragraph:

Features of this coder include:

- --Two speech modes: voiced and unvoiced
- --Unvoiced mode uses stochastic excitation codebook
- --Voiced mode uses sparse pulse codebook
- --20 ms frame size, 10 ms subframe size, 2.5 ms LPC subframe size
- --Perceptual weighting applied in codebook search